

Associate Lecturer Cristina BURLĂCIOIU, PhD

E-mail: cristina.burlacioiu@asebuss.ro

ASEBUSS – The Institute for Business Administration in Bucharest

Professor Cristina BOBOC, PhD

E-mail: cristina.boboc@csie.ase.ro

Department of Statistics and Econometrics

The Bucharest University of Economic Studies

Bogdan MIREA

E-mail: mireabogdan17@stud.ase.ro

The Bucharest University of Economic Studies

Ioana DRAGNE

E-mail: ioana.dragne8@gmail.com

Advanced Digital Imaging Techniques at Faculty of Electronics,

Telecommunications and Information Technology

University POLITEHNICA of Bucharest

TEXT MINING IN BUSINESS. A STUDY OF ROMANIAN CLIENT'S PERCEPTION WITH RESPECT TO USING TELECOMMUNICATION AND ENERGY APPS

***Abstract.** New technologies have enabled companies to create new and contemporary designs and, at the same time, increase the efficiency of their operations. The Internet has amplified not only the e-commerce with products, but also with services. Paying and changing contracts online in only a few minutes for telecommunication and energy services are only two of the advantages of online APPs. The objective of this study is to show how to use and analyze the review data from customers on online applications or forums, providing very useful insights that could be used to improve company-customer relationships and gain business benefits. The data were collected by using the web scrapping method, obtaining a large amount of text. For the overall understanding of this large amount of text, sentiment analysis was used. Then, to classify the reviews into one of three classes: negative, positive or neutral feeling, five methods were used and compared: logistic regression, decision tree method, K Nearest Neighbors method, SVM algorithm, Naïve Bayes.*

***Keywords:** machine learning, web scrapping, online APP`s, classification, sentiment analysis.*

JEL Classification: C45, C82

1. Introduction

Artificial intelligence plays an increasingly important role in data processing due to the increase in the volume of information available in the virtual environment. Businesses use data analytics algorithms to predict trends, identify opportunities, and understand consumer behaviour before they decide what to buy or use. Thus, companies that implement artificial intelligence systems form their decision-making models based on complex analyses. The main methods used in the category of technologies that analyse written language (text mining) are topic identification, information extraction, natural language processing, sentiment analysis, and various visualisation methods, all applied to solve a concrete business need in a fast and exact manner.

The paper, following the stages of a text analysis, from the building of database, the pre-processing of the obtained data and the descriptive analysis, to the sentiment analysis of the reviews and the use of some classification models provides an example of the processing and analysis of the reviews on a platform of applications from the main Romanian telecom and energy companies. First, we analysed the differences between the telecommunications and energy industries in the perception of Romanian consumers towards the APPs used in these two industries. Second, the telecommunications and energy industries are compared in terms of the model's performance used to classify the feelings of the customers. In order to implement the proposed analysis methodology, the Python programming language was used.

2. Literature review

One of the most important aspects in a successful company is innovation. Steps towards change often come with doubts and mistrust, both from employees, who may not understand business directions, but also from existing or potential customers, who may see any change as a less good thing when they choose to use a product or service. That is why companies seek to understand users' needs and come up with the best possible solutions to meet them, in a proactive way. In recent years, the trend of implementing data analysis algorithms in Big Data has grown considerably (Angelopoulos & Kontakou, 2021). This new trend is part of the use of new technologies, Artificial Intelligence (AI), the use of IoT (Internet of Things) technology, Cloud work and storage, or Smart technologies.

Topic identification, sentiment extraction, summarisation, and even translation are solutions specific to Artificial Intelligence. Thus, the intervention of a person who would have had to go through all the text-type information collected from customers is no longer necessary. Algorithms show their usefulness not only through the speed of execution but also through increased efficiency.

We can think of reviews as a more advanced method of social listening, allowing us to look behind the curtain and find out exactly what customers want in their own words. This information is extremely useful to detect and correct any operational inefficiencies (O'Brien, 2017).

Abstract feelings and beliefs, such as preferences, personal estimation, or level of appreciation, have a huge impact on human behaviour. Because expressing judgments of others is naturally embedded in each individual and presents us as "social beings", the choices we make can be greatly influenced by others' perceptions of the universe. Social media platforms such as Twitter, Facebook, and YouTube have exploded in popularity in recent years (Mehta & Pandya, 2020).

There are many solutions documented in research on the use of text mining techniques in online reviews in various fields. In the tourism industry, accommodation capabilities are an essential part of attracting tourists (Jia, 2020). Service providers extensively use online reviews for tourist destinations marketing. Positive and negative comments give consumers a clear picture of the tourist destination and provide service providers with support to understand tourists' needs (Guerreiro, 2017). Topic identification, sentiment analysis, and regression techniques help to understand customer opinion about restaurants (Pezenka&Weismayer, 2020). Online consumer reviews are used by the airline industry to outperform the competition. The main objective of the airline industry is customer satisfaction. Sentiment analysis and topic modelling are widely used approaches in the airline industry to assess overall consumer happiness (Kim&Lee, 2019).

Research on sentiment analysis in the Romanian language is not numerous. Of these, one analysis uses a database of 1000 movie reviews, of which 500 are categorised as positive reviews and 500 as negative reviews. The algorithms used were Random Forest, KNN, Naive Bayes and SVM, the best precision and recall indicators, of 81.8%, had the Naive Bayes algorithm used together with the SentiWordNet lexicon, which classified the text from Romanian translated into English with the Microsoft API Translator (Russu et al., 2014). Other works use BERT, a natural language analysis method developed by Google with 93% accuracy on the validation data set, and synthesise the results into a model that can later be used for sentiment prediction (Cioban, 2021; Masala &al., 2020). Another approach to text analysis in Romanian language is presented in a work that aims to build a sentiment analysis model used for labelling centralised responses following a psychological survey in categories of thoughts, emotions, and behaviours (Neagu et al., 2022; Lupe & Briciu, 2019).

3. Research methodology

The process of making strategic organisational decisions in today's complex world is a dynamic process characterised by uncertainty.

The adaptation and implementation of analysis models based on intelligent algorithms in current business systems has grown in recent years. The pandemic period has accelerated the rate of adoption of digital solutions, with 52% of companies interested in increasing the involvement of AI in business processes (PwC, 2021).

NLP (Natural Language Processing) – Natural Language Processing or computational linguistics is a process by which a text is processed in a way

recognised by a computer. Thus, the text can be analysed using modern analysis techniques such as automatic translation, identifying the feeling of a text, or identifying the topic. The main goal of NLP is to remove the understanding barrier between computers and humans, text data is also taken from speech, through processes of transforming audio recordings into text, which then become matrices of binary indicators.

Sentiment analysis (also known as “opinion mining”) is a procedure that belongs to Natural Language Processing and aims to classify text into 3 categories: negative, positive, or neutral text. The technique is used especially by the business environment that wants to have an overview of the customers' opinions on the products they use. An important asset of sentiment analysis is the exponential increase in the amount of data available online, as more and more people use the virtual environment to express their opinions.

Analysis in several languages, analysis based on different aspects, or emotion detection are just a few branches that have developed from the need for an exploration as suitable as possible for the type of data and business requirements encountered.

Learning algorithms based on sentiment analysis are the next step in the classification of reviews collected from the online environment. Supervised learning is an approach to train algorithms with input data for the independent variables and output data for the dependent variable. The supervised learning algorithm performs a function that can be used in the future to estimate the values of the target variable with the help of new values for the independent variables.

The use of Natural Language Processing (NLP) has shown its usefulness in several business applications, from virtual customer assistants to bots able to answer questions from users, systems for generating written reports, or topic classification models.

Sentiment analysis is a tool powered by AI algorithms and implemented through NLP that provides actionable market research by analysing comments, customer reviews, social media mentions, and other digital footprints. NLP is implemented using a sentiment analysis model that converts and classifies words and phrases into actionable marketing information. The algorithm identifies frequently used phrases and classifies them as positive, neutral, or negative. This information is transformed into meaningful knowledge that is used to generate marketing campaigns, identify the most appreciated functionalities, or frequent bugs.

4. Data description

Paying and changing contracts online in only a few minutes for the TELCO and ENERGY services are only two of the advantages of online applications. This study has attempted to see the differences in the perception of Romanian consumers towards the APPs used in the two main industries offering essential services: telecommunications and energy.

The data used in this paper consists of a set of mobile applications reviews collected from *play.google.com* using a web scrapping package implemented in Python. We use the reviews of the main telecommunications (Orange, Vodafone,

Telekom, Digi) and energy (Enel, Electrica, CEZ, E.ON, engie) companies’ mobile applications in Romania: My Orange Romania, My Vodafone, MyAccount Telekom, Digi Romania, MyEnel Romania, MyElectrica, My CEZ, E.ON Myline, and engie mobile.

During 2020-2021, we collected a total of 53,720 records. The dataset contains mainly Romanian text, with some comments also in English. Each review contains:

- 1) the name of the application,
- 2) the date the review was published,
- 3) the score given by the users (from 1 to 5 stars),
- 4) the review’s text.

The dataset was split in two subsets – telco and energy – and then comparative studies were conducted on them.

Table 1. Overview of the scores

Industry	Telco market My Orange Romania, My Vodafone, MyAccount Telekom, Digi Romania					Energy market MyEnel Romania, MyElectrica, My CEZ, E.ON Myline, Engie mobile				
	Negative (0)		Neutral (1)	Positive (2)		Negative (0)		Neutral (1)	Positive (2)	
Score	1	2	3	4	5	1	2	3	4	5
No. reviews	23795	3494	3471	6473	51373	8702	1097	834	870	4337
[%]	27%	4%	4%	7%	58%	55%	7%	5%	6%	27%

The results show that for the telecommunications market, most scores are positive (65%), while for the energy apps, 62% of the scores are negative.

The distribution of the length of reviews shows us that most comments are between 0-19 characters long, with the number of comments decreasing as the number of characters increases. We notice that for each category, the most frequent words remain similar. The peculiarity derived from reviews analysis is that users give suggestions when they assign low marks, but also when they leave a good opinion about the product.

Following the analysis of the words and their frequency, for each of the three defined classes, we find a model that must be taken into account:

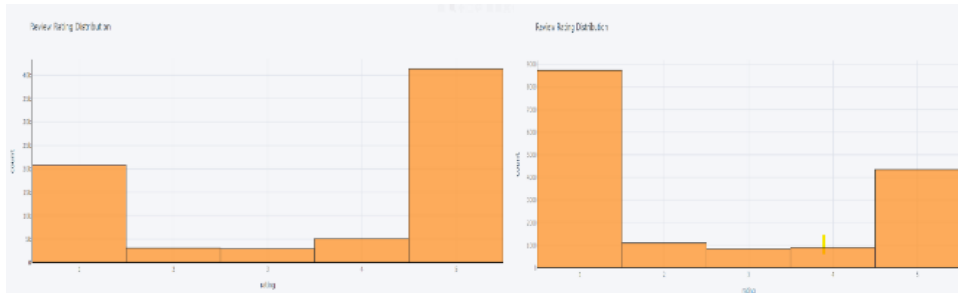


Figure 1. Rating distribution for (right) TELCO and (left) ENERGY mobile apps reviews

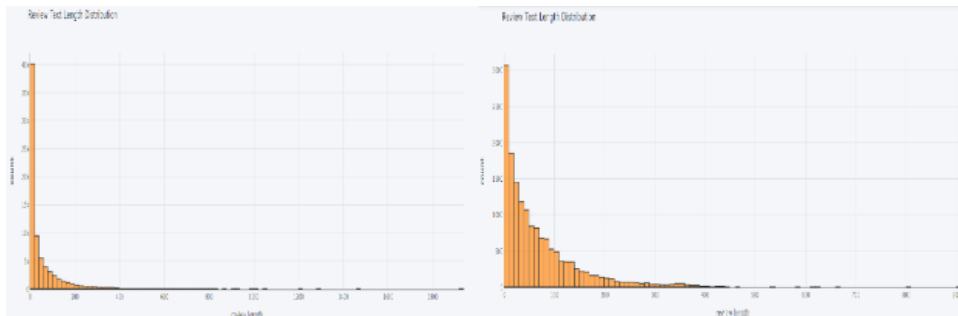


Figure 2. Reviews length distribution for (right) TELCO and (left) ENERGY mobile apps reviews

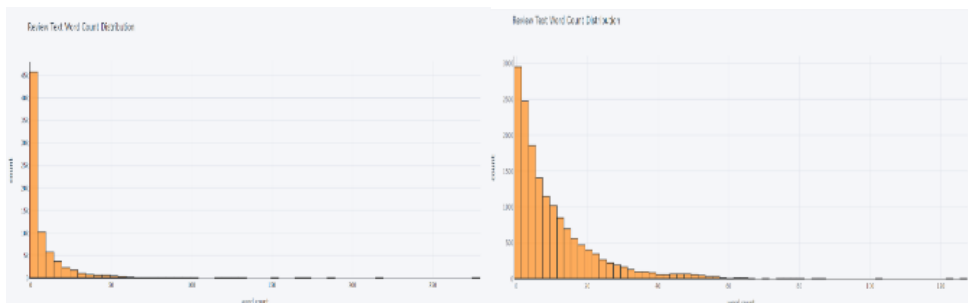


Figure 3. Words count distribution for (right) TELCO and (left) ENERGY mobile apps reviews

In the case of positive sentiment, words like **"useful"** and **"helpful"** show satisfaction with the app. Other phrases such as **"simple"** and **"easy to use"** show us that the way the application was created is suitable for users, although some comments classified with positive sentiment contain suggestions for improvement that should not be ignored, **"can't log in"** and **"can't connect"** showing that improvements are needed on the connectivity side if we want to keep the positive customer impression.

Text Mining In Business. A Study of Romanian Client's Perception with Respect to Using Telecommunication and Energy Apps

In the case of negative sentiment, the main complaints are also related to connectivity and logging into the app, which places this issue as the main task to accomplish in order to reduce negative sentiment among customers and keep positive the sentiment of the customers who still do not consider connectivity a serious problem.

Neutral sentiments group a very important class of users, as they can change their opinion much faster. Here, in addition to the suggestions from the other two categories, we identified problems with paying the bill and entering the password. Solving these problems can attract a positive feeling for customers and increase their satisfaction with the services provided by the company.

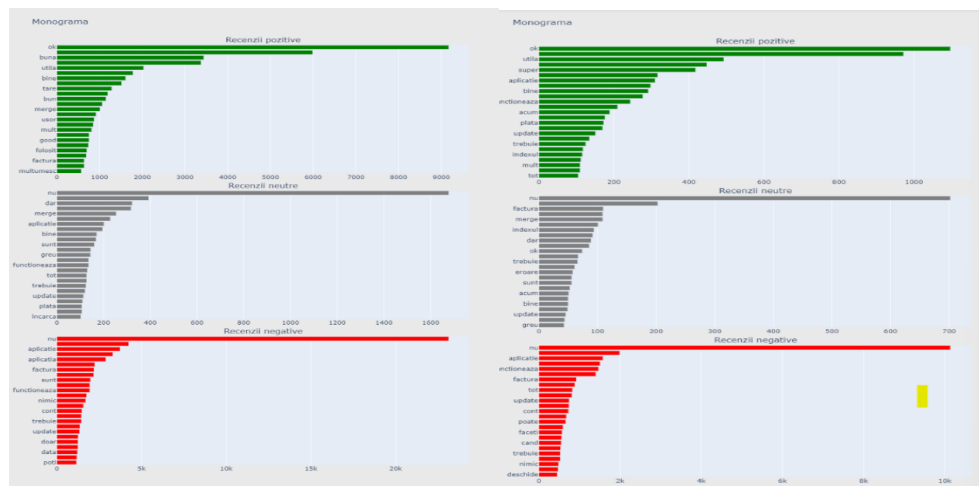


Figure 4. Monogram for (left) telecommunication and (right) energy mobile apps

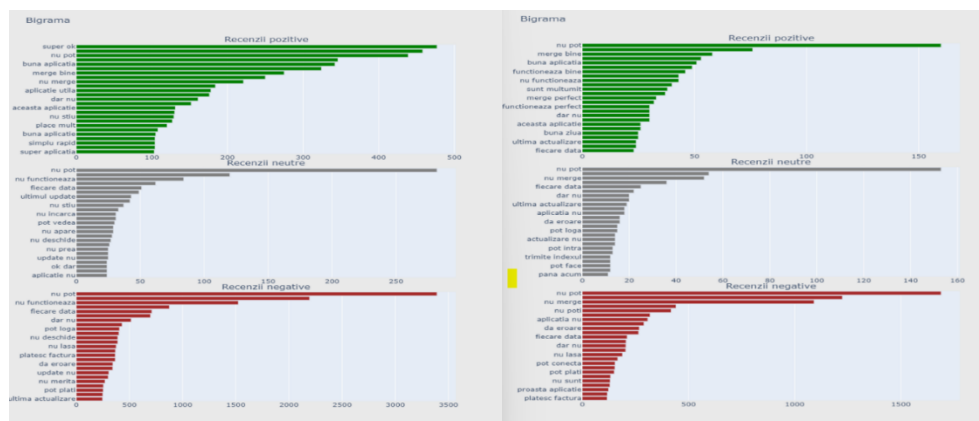


Figure 5. Bigram for (left) telecommunication and (right) energy mobile apps reviews

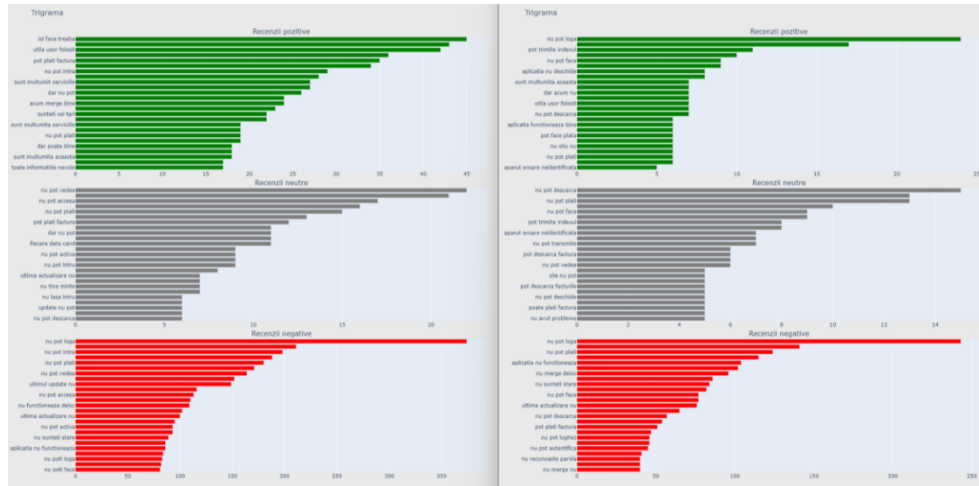


Figure 6. Trigram for (left) telecommunication and (right) energy mobile apps reviews

5. Analysing reviews using ranking algorithms

5.1 Text vectorisation

An important step was preprocessing the data so we could feed it as the input data for the proposed models. We started by removing/replacing diacritics and unwanted special characters with their corresponding letter, making the text lowercase, removing punctuation and numbers. Then, a custom stopwords list was set. As the reviews were written in both Romanian and English, the stopwords list was defined to cover the commonly used words that appear to be of little value for both languages. By removing these words from the reviews, we can focus on the relevant words instead.

The Porter stemming algorithm was then applied to reduce the words to their root dictionary form, known as the lemma (e.g., choosing – chose).

To determine the relevance of a word in a document in a series of documents, the TF-IDF or "Term Frequency – Inverse Data Frequency" method is applied. The method is used in various ways, the most important being in text analysis, showing its usefulness in word classification in specific Machine Learning algorithms for Natural Language Processing. The score is calculated using two indicators: the frequency of occurrence of a word in a document and the inverse frequency of occurrence of the word in the set of documents (Bijoyan&Chakraborty, 2018).

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D),$$

$$\text{where: } TF(t, d) = \log(1 + \text{freq}(t, d)),$$

$$IDF(t, D) = \log\left(\frac{N}{\text{count}(d \in D: t \in d)}\right)$$

The goal of the process of sentiment classification in our case is given a review, classify it as a positive/ neutral/ negative one. This is a supervised learning case and training data for this classification requires data labels/ ground truth tags. Manually labelling each review was not a solution given the amount of data we collected; thus, we used *polyglot* to automatically tag our dataset.

Table 2. Reviews' polarity distribution

Telco			Energy		
Negative sentiment	Neutral sentiment	Positive sentiment	Negative sentiment	Neutral sentiment	Positive sentiment
6595	70194	11817	1122	13958	742
7%	79%	13%	7%	88%	5%

After labelling the reviews, we can already observe the difference between the rating distribution and the sentiment distribution: more than 70% of the reviews seem to have a neutral message, but only 5% of the ratings are neutral (3 stars). This denotes that either the user wrote an ironic text, either the user rated the application with a wrong score; e.g., “best application” with 1 star review could mean that the application is very bad (1 star) and the comment is purely ironic, either the application really is the best, but the rating was incorrectly chosen. Note that our dataset does not contain one language only text and polyglot is not sensitive enough to language changes (e.g., “super” or “ok” are used in both Romanian and English and for these cases we cannot rely on the labelling resulted from polyglot). We will use as labels the categories set according to the rating.

The "feeling" variable has to be numerically coded for future analyses: 0 – negative feeling, 2 – positive feeling, and 1 – neutral feeling. We observe a big difference between the number of observations in each category (in category 2 there are 34198 observations, in category 0 there are 17556 observations; and in category 1 there are 1966 observations). Through the SMOTE method, the 3 classes are balanced. SMOTE replicates text classes by selecting one or more classes based on the k-nearest neighbors’ algorithm.

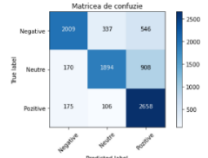
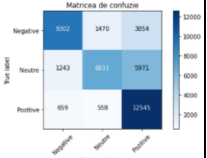
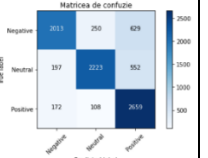
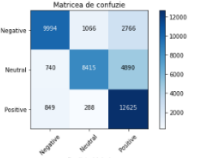

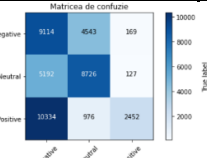
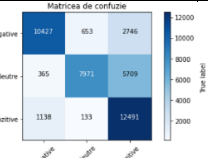
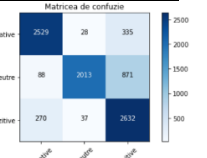
Thus, the text is transformed into numbers that can be processed by supervised learning algorithms, a procedure also known as text vectorisation. The data set was divided into two: in train and test (70:30).

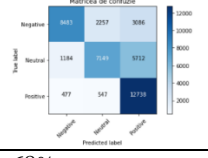
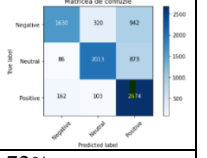
For sentiment classification, the following methods were used:

- Logistic regression is a statistical model that offers the possibility of predicting a discrete binary or multinomial variable by means of several discrete and continuous variables. Logistic regression is useful for our case as its distribution is an S-curved distribution (cumulative density function) that is similar to the standard normal distribution and constrains the estimated probabilities to be in the interval (0,1). We used the classifier implemented by the Python scikit-learn library, *LogisticRegression*.
- A decision tree is a decision support tool that uses a tree-like model of choice and its possible consequences. It is a presentation method of an algorithm composed of conditional control instructions. The decision tree classifier is used from scikit-learn library using the *DecisionTreeClassifier* instance.
- The K Nearest Neighbors method is a non-parametric method used for classification and regression. The KNN classifier compares the labeled (classified) observations in a train data set with the K closest unlabeled observations in the test set to see if the unlabeled observations share the same label as the labeled ones.
- The SVM algorithm is a supervised learning algorithm aimed at classification. Being a binary classification technique, it uses the input data set to form an optimal hyperplane in an n-dimensional space. The hyperplane is used to classify new input data. SVM algorithms are used to classify data in a two-dimensional plane or a multidimensional hyperplane. The multidimensional hyperplane uses "centers" to classify data. There is a clear distinction between classified data points. For that, there should be a maximum distance or the hyperplane should allow a maximum margin for the data set. SVM is known to be an accurate classifier for text documents.
- Naïve Bayes is a widely used model in classification, however, it assumes that all variables are conditionally independent, but it is hard to achieve it in text analysis. The words in a document are not independent among them.

Applying these models, the results obtained are presented in Table 3. The overall best results are obtained for the SVC and Decision Trees models, and the worst result is obtained for the kNN model. For the energy market, the results are better. One reason could be the smaller dataset and the better data cleaning, fewer words in different languages, while on the telco there are many reviews in English, which will lead to a weaker processing. The execution time for SVC is higher than for the other algorithms. Considering the results versus the execution time, decision trees could be the recommended method to choose.

Table 3. Models evaluation

Model	M1. Logistic regression		M2. Decision trees	
Industry	Telco	Energy	Telco	Energy
Confusion matrix				
Accuracy	69%	75%	75%	78%
Precision	73%	77%	78%	80%
Recall	69%	75%	75%	78%
F1-score	68%	74%	75%	78%
Model	M3. kNN		M4. SVM	
Industry	Telco	Energy	Telco	Energy
Confusion matrix				
Accuracy	49%	46%	74%	81%
Precision	62%	50%	79%	84%
Recall	49%	46%	74%	81%
F1-score	46%	44%	74%	82%

Model	M5. Naïve Bayes	
Industry	Telco	Energy
Confusion matrix		
Accuracy	68%	72%
Precision	72%	76%
Recall	68%	72%
F1-score	67%	72%

Note:

Tested models: Logistic Regression, kNN, SVC, Native Bayes

Hyperparameters tuning method: Grid Search

Train-test split: train - 70%, test - 30%

Metrics: Accuracy, Precision, Recall, F1 score

6. Conclusions

In this paper, we provide a comparative study between existing techniques associated with sentiment analysis including Machine Learning approaches, applied on customer reviews data.

For data collection, we used web scraping to extract the reviews from Romanian telecom and energy mobile applications. After the feature selection, data cleaning was an important and challenging step by having to remove or transform lots of special characters and misspellings. TF-IDF text vectorisation was used to transform input data before feeding it to various Machine Learning models, while SMOTE was implemented to solve the challenge of working with an imbalanced dataset.

Next, we chose five of the most widely used Machine Learning models for sentiment analysis: logistic regression, decision tree, K Nearest Neighbors, SVM, Naïve Bayes. As the algorithms' effectiveness depends on the characteristics of the dataset, we split the dataset in two – telecom and energy – and conducted some experiments on both subsets to analyse and evaluate the models, in order to cover a greater domain variety. Research results show that the decision tree and SVM have the best accuracy for all 3 analysed classes, but in terms of execution time, the decision tree outperforms SVM. For telco, an average accuracy of 75% was obtained for the two models; for energy, the results increased to 79.5% on average. Most of the negative reviews are related to faulty in app functionalities, while positive reviews praise the ease of use, simplicity, or swiftness of the application.

Sentiment analysis is a powerful tool when it comes to gaining insight into users' reviews, since it allows companies to better understand the clients' needs and which functionalities and services they should add or improve. The study shows that almost 70% of the neutral reviews have a confusing rating associated. Sentiment classification could be used to better understand the users' opinion, as the review alone could provide a false impression when the rating is not taken into consideration and vice-versa.

Supervised machine learning techniques present a good balance between accuracy and runtime for sentiment classification. Future work could be extending the study to new models and different methodologies for text vectorisation, so we can improve both accuracy and the capacity to adapt to various languages and levels of irony.

REFERENCES

- [1] **Angelopoulos, M. K., Kontakou, C. (2021), *Data Analytics and Customer Experience in Industry: An Option*. *International Journal of Advances in Engineering and Management*, 3(3), 1294-1300, www.ijaem.net, ISSN: 2395-5252, DOI: 10.35629/5252-030312941300;**
- [2] **Bijoyan, D., Chakraborty, S. (2018), *An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation*. *ArXiv*, [abs/1806.06407](https://arxiv.org/abs/1806.06407), <https://doi.org/10.48550/arXiv.1806.06407>, <https://arxiv.org/ftp/arxiv/papers/1806/1806.06407.pdf>;**
- [3] **O'Brien, M. (2017, Noiembrie 9), *Reviews and refinement: Using customer reviews to identify areas for improvement*. Taken from Clickz: <https://www.clickz.com/reviews-and-refinement-using-customer-reviews-to-identify-areas-for-improvement/>;**
- [4] **Cioban, Ș. (2021), *Cross-Domain Sentiment Analysis of the Natural Romanian Language*. In: Jallouli, R., Bach Tobji, M.A., Mcheick, H., Piho, G. (eds) *Digital Economy. Emerging Technologies and Business Innovation. ICDEc 2021*. Lecture Notes in Business Information Processing, vol 431, Springer, Cham. https://doi.org/10.1007/978-3-030-92909-1_11;**
- [5] **Guerreiro, J., Moro, S. (2017), *Are Yelp's tips helpful in building influential consumers?*. *Tourism Management Perspectives*, 24, 151-154, DOI:10.1016/j.tmp.2017.08.006;**
- [6] **Jia, S. (2020), *Motivation and satisfaction of Chinese and U.S. tourists in restaurants: A cross-cultural text mining of online reviews*. *Tourism Management*, 78, 104071 ref.111, DOI: 10.1016/j.tourman.2019.104071;**
- [7] **Kim, S., Lee, W.S. (2019), *Network text analysis of medical tourism in newspapers using text mining: The South Korea case*. *Tourism Management Perspectives*, 31, 332-339, DOI:10.1016/j.tmp.2019.05.010;**
- [8] **Lupea, M., Briciu, A. (2019), *Studying emotions in Romanian words using formal concept analysis*. *Computer Speech & Language*, 57, <https://doi.org/10.1016/j.csl.2019.02.005>;**
- [9] **Masala, M., Ruseti, S., Dascalu, M. (2020), *RoBERT – A Romanian BERT Model*. 6626-6637, DOI: 10.18653/v1/2020.coling-main.581;**

- [10] **Mehta, P., Pandya, D. (2020), A Review On Sentiment Analysis Methodologies.** *International Journal of Scientific and Technology Research*, 9, 601-607, ISSN 2277-8616, <http://www.ijstr.org/final-print/feb2020/A-Review-On-Sentiment-Analysis-Methodologies-Practices-And-Applications.pdf>;
- [11] **Neagu, D.C., Rus, A.B., Grec, M., Boroianu, M.A., Bogdan, N., Gal, A. (2022), Towards Sentiment Analysis for Romanian Twitter Content.** *Algorithms* 2022, 15, 357. <https://doi.org/10.3390/a15100357>;
- [12] **Pezenka, I., Weismayer, C. (2020), Which factors influence locals' and visitors' overall restaurant evaluations?.** *International Journal of Contemporary Hospitality Management*, 32(9), 2793-2812, DOI:10.1108/IJCHM-09-2019-0796;
- [13] **PwC. (2021), AI Predictions 2021.** Taken from [pwc.com: https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html](https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html);
- [14] **Russu, R. M., Vlad, O., Dinsoreanu, M., Potolea, R. (2014), An Opinion Mining Approach For Romanian Language.** *Proceedings of IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 43-46, https://keg.utcluj.ro/papers_2014.php.